Decision Support System (DSS) for Fraud Detection in Health Insurance Claims using Genetic Support Vector Machines (GSVMS)

ANU VICTOR, SP CHOKKALINGAM
Department of Computer Science and Engineering,
School of Computing,
Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science & Technology, Chennai,
INDIA

Abstract: - Fraud in health insurance leads to huge financial losses and claim processing systems inefficiencies. In this paper, the proposed system to be used is a hybrid Genetic Support Vector Machine (GSVM) model of a Decision Support System that will be effective in detecting fraud. FIC-GSVM method combines Genetic Algorithms to select the most optimal features and Support Vector Machines to classify features. The model is tested with the National Health Insurance Scheme (NHIS) data on Ghana, which is a set of structured and coded healthcare claims. Experimental findings indicate that the proposed approach has a 90.5% accuracy, as well as high precision, recall, and reduced processing time as compared to the current approaches. The system facilitates automated, scalable, and near real-time fraud detection, which can be an effective solution to healthcare insurance analytics.

Key-Words: - Health Insurance Fraud Detection, Decision Support System (DSS), Genetic Algorithm (GA), Support Vector Machine (SVM), Genetic Support Vector Machine (GSVM), Feature Selection, Machine Learning, Healthcare Analytics.

Received: July 19, 2025. Revised: December 11, 2025. Accepted: March 19, 2026. Published: June 23, 2026.

1 Introduction

Healthcare fraud negatively impacts insurance companies' revenues, [1]. Consumers and intermediaries misuse fake or misleading information to get fraudulent payments. The health insurance business has lost customers' confidence, and adjustment costs have increased, [2]. Manual and inefficient rule-based fraud detection systems are used. Hidden patterns in complex and huge data sets cannot be revealed by these methods, [3]. Innovations that can accurately and automatically detect fraud are needed now. In the last decade, machine learning surfaced as an efficient way to automate fraud detection processes, [4]. Because of the speed and accuracy that it has in labeling datasets of considerable size, it works to find evolving, sophisticated patterns that may elude human recognition. One approach is to create a hybrid framework called GSVMs, which combines GA with SVM, [5].

GSVMs harness the advantages of GA and SVM, combining the ability for feature selection from GA with the strong classification ability of SVM, [6]. This hybrid framework allows for good classification while improving accuracy and reducing human effort, both existing downsides of traditional methods, [7]. However, GSVM encountered challenges with noisy data or in spatially-structured datasets where there are more features than samples, [8]. Nonetheless, GSVMs have good performance on structured datasets that provide labeled data for health insurance claims, [9]. This study will use data obtained from Ghana's National Health Insurance Scheme and apply the GSVM framework to it. The goal of the study is to accurately categorize claims and quickly identify fraud, [10]. The proposed approach will require fewer human efforts to evaluate claims, complete the process faster, and improve fraud detection performance.

1.1 Motivation

Global health insurance fraud continues to rise, causing financial losses, higher premiums, and public distrust in healthcare financing systems. Upcoding, duplicate billing, ghost claims, and exaggerated service costs cause systemic inefficiencies and skew actuarial risk assessment. Nonlinear and changing fraud patterns are difficult to identify using rule-based audits and human verification, which are time-consuming and resource-intensive. These methods frequently have large false-positive rates and slow detection cycles. The proposed approach uses Genetic Support

Vector Machines (GSVMs) to automate detection and improve operational efficiency. The system uses evolutionary feature optimization and margin-based classification to enhance prediction discrimination, minimize inappropriate payments, and allow scalable healthcare insurance fraud monitoring.

1.2 Problem Statement

Health insurance fraud, especially fraudulent financial aid claims, causes enormous economic losses and administrative strain. Fraud-related abnormalities in Ghanaian structured hospital datasets include anomalous billing frequencies, exaggerated treatment costs, inconsistent diagnosis–procedure pairings, and repetitive provider–beneficiary contacts. In high-dimensional, noisy data settings, human audits and static threshold criteria hamper scalability and flexibility of traditional detection systems. GSVM offers a solid classification framework, however healthcare claim data's repetitive characteristics, feature correlation, and class imbalance provide issues. An automated fraud detecting system using genetic feature selection for dimensionality reduction, optimal SVM hyperparameter tuning for margin stability, and structured data preparation for noise mitigation is developed in this paper. Following page 4, further discussion illustrates the proposed system's operating procedure for clarity. A claim with abnormally high laboratory costs and frequent short-interval admissions is turned into a feature vector and analyzed by the GSVM decision boundary to calculate fraud likelihood. How the evolutionary algorithm repeatedly picks discriminative information like claim amount variation and provider frequency index improves classification robustness is explained. This revised description clarifies methodology, improves contextual relevance, and sheds light on healthcare insurance fraud detection system deployment.

1.3 Contributions

The major contributions of this paper are;

- This study offers a Hybrid FIC-GSVM capable of accurately identifying fraudulent health insurance claims.
- The improved claim classification accuracy, reduced time to process claims, and reduction in the time a human is required to assist in the detection of fraud offer improvements over traditional rule-based methods through the use of GSVMs.
- The framework is evaluated on real-world NHIS health insurance data from Ghana to determine its ability to identify unusual patterns and deviations in real-world health insurance claims.

The rest of this paper is arranged as follows: Section 2 evaluates research on financial losses and claim processing system trust. FIC-GSVM is suggested in Section 3. Section 4 compares this technique with others. Finally, Section 5 discusses a future study.

2 Related Works

Recent advances in machine learning have significantly increased the ability of financial and insurance organizations to detect fraud. Researchers have explored hybrid models, which combine clustering, optimization, and classification techniques, as a possible way to achieve greater accuracy and efficiency. Because they are such reliable classifiers, SVMs are well established in classification applications. Current studies are focused on designing smart systems that can learn and adapt to shifting patterns of fraud.

The proposed hybrid model for fraud detection in this paper comes from the combination of the Firefly Optimization Algorithm (FOA) with SVM, [11]. With the FOA for optimal feature selection, this supports the SVM classification accuracy for financial data. Incorporating both FOA and SVM led to a positive increase in detection accuracy and a decrease in computational complexity. The experimental results of (FOA-SVM) show a significant increase in fraud detection compared to more traditional machine learning models.

This study offers a two-stage hybrid model (T-SHM) that uses K-Means clustering on the data, which is now partitioned into groups, and then uses a support vector machine to classify and separate the groups at risk of fraud, [12]. In the context of insurance claims data, K-Means can uncover potential groups or patterns, and SVM provides a method of classifying them for the purpose of fraud mitigation. By reducing the noise in the data and improving decision boundaries, the process improves detection rates and accuracy.

The authors describe a new method to detect fraud using an iterative SVM and a map-reduce framework (FD-SVM-MRF). This method is essential for the scalable processing of large healthcare insurance datasets, [13]. The system trains the SVM classifier using data in an iterative manner, where the data chunks are processed in parallel in a distributed system to provide significant performance increases and reduced

training time. The authors showed that their system was able to successfully detect fraudulent healthcare claims in a big data environment.

The repetition of this research paper on the list illustrates how FOA combines with SVM once again. By selecting the most useful features found in the financial data, the FOA improves SVM classification accuracy, [14]. The hybrid model, which encodes more information than a standalone model, can help prevent overfitting by eliminating noise, reducing the number of attributes for the model, and focusing on attributes and noise. A high-dimensional financial dataset found in the real world is ideal to use, due to its powerful results.

The goal of this research is to improve insurance analytics with feature selection and machine learning models (FS-MLm). The authors use several machine learning algorithms that include feature selection methods to find which variables significantly affect claim behavior, [15]. This is aimed at improving the accuracy of predictions and obtaining actionable insights from insurance databases. Improved fraud detection and risk estimation are achieved with the proposed methodology.

The researchers suggested the Blockchain-assisted healthcare insurance fraud detection framework using ensemble learning. Decentralized blockchain technology protects patient and healthcare data. The author compare ensemble learning approaches like bagging classification and stacking to standard Machine Learning Algorithms (MLAs) to evaluate our methodology, [16]. The novel technique integrates in-patient, out-patient, and beneficiary data, going beyond established methodologies. This extensive integration makes our solution more applicable in the real world and provides a holistic view of healthcare insurance claim fraud detection. Accuracy, precision, recall, ROC, 11-score, and confusion matrix are evaluated.

The authors proposed the Machine Learning for Fraud Detection and Error Prevention in Health Insurance Claims. Healthcare insurance handles millions of claims every day, making it a hotbed for fraud and mistakes. In recent years, health insurance prices have skyrocketed due to insurance companies' payment irregularities during claims processing. Our machine learning technology detects fraud and prevents mistakes, [17]. Deep learning, supervised and unsupervised learning, and natural language processing examine massive information to find claims data patterns, abnormalities, and contradictions. Reprocessing claims commonly follows insurance company payment errors. Rework involves reprocessing claims. Machine Learning enhances accuracy, cost, claims processing, and customer happiness. Future Machine Learning will enable real-time decision-making and industrial cooperation.

Health insurance fraud detection using multi-channel heterogeneous graph structure learning was suggested in [18]. Multi-channel Heterogeneous Graph Structured Learning (MHGSL) was recommended for health insurance fraud detection. For health insurance data from patients, departments, and drugs, MHGSL employs graph structure learning to extract topological structure, features, and semantic information to generate various graphs that represent data variety and complexity. Multi-channel information transfer and feature fusion uncover health insurance data anomalies using heterogeneous graph neural networks and graph convolutional neural networks. According to extensive studies on actual health insurance data, MHGSL identifies possible fraud better than previous approaches and can rapidly and reliably identify fraudulent patients to avert health insurance fund loss. Experimental results suggest MHGSL multi-channel heterogeneous graph structure learning may help detect health insurance fraud.

The Bayesian Belief Network to detect healthcare fraud. The author demonstrates how fraud detection systems may benefit from Bayesian Belief Network (BBN), a graphical network model that harnesses the relational nature of transaction data and has superior interpretability than many competing machine learning approaches, [19]. Testing with a real-world insurance claims dataset shows that BBN model performance is equivalent to logistic regression and random forest. The best-performing BBN has a 0.15 F1 fraud class score. Based on common sense qualitative assessment, auditors and investigators may evaluate the suggested BBN model.

The paper introduced the advanced blockchain-based hyperledger fabric solution for tracing fraudulent claims in the healthcare industry. For newly admitted patients, the model proposes an insurance policy, but for current patients, it speeds up transaction processing and secures payment settlement using a private blockchain [20]. These two technologies might change the future. In this study, blockchain and ML techniques like Support Vector Machine (SVM) and Random Forest Regression will distinguish fraudulent and legal medical records to recommend personalized policies, streamline claim processing, and protect sensitive patient and vital insurance data. The goal is data openness and patient-centricity. An

integrated framework creates a safe, flexible, and efficient environment that surpasses old approaches, enabling the future of healthcare and insurance.

Using SVM with FOA, K-Means, or the map-reduce architecture, the reviewed articles improve the accuracy of fraud detection. Fraud classification brings attention to many facets of research, including feature selection, scalability, and automation. Their performance has been verified with real-world datasets and confirmed their system performance in the areas of finance and insurance, as examined and given in Table 1 (Appendix). All of these methods aid in enhancing the ease and accuracy of fraud detection mentioned in Table 1 (Appendix).

3 Proposed System

Healthcare insurance fraud in general incurs significant financial waste and operational inefficiency. Traditional fraud detection technologies, which are typically slow and not well-equipped to analyze complex and massive datasets, can be a significant burden. In contrast, using machine learning allows users to automate their data patterns analysis and discover previously undiscovered patterns in organized and semi-structured lists. The proposed outcomes for the hybrid system are to automate the claims processing function, improve accuracy, and reduce labor costs.

3.1 Overview of FIC-GSVM System

Genetic Algorithms and SVM are used in the suggested system to make a hybrid model called FIC-GSVM. This model can help find health insurance fraud. The system acts as a DSS to help sort health insurance claims by automating the process, speeding up the time it takes to process claims, lowering rewards for false claims, and lowering the number of mistakes made by people.
However, the system is fully automated to check if an insurance claim is real or fake using the best parts of GA and SVM in Figure 1 (Appendix). Rule-based systems have problems, such as being slow and not always correct. This hybrid approach is meant to avoid those problems. The GSVM system may be employed in real life since it learns from an organized and labeled dataset. The NHIS is one example of a labeled dataset in Ghana's health care system. In conclusion, this summary shows how FIC-GSVM can be a smart and affordable way to find health insurance fraud and change the way claims are currently being handled.

Algorithm 1: FIC-GSVM-based algorithm
function $FIC\_GSVM\_FraudDetection(claim\_dataset)$:
// Step 1: Data Preprocessing $leaned\_data = PreprocessData(claim\_dataset)$

// Step 2: Genetic Feature Selection $optimal\_features = GeneticFeatureSelector(cleaned\_data)$

// Step 3: Train Support Vector Machine $svm\_model = TrainSVM(cleaned\_data[optimal\_features])$

// Step 4: Classify Insurance Claims $classified\_results = svm\_model.Predict(claim\_dataset[optimal\_features])$

// Step 5: Evaluate Model Performance $evaluation\_metrics = EvaluateModel(classified\_results)$
$return\ classified_{results}, evaluation_{metrics}$

This solution uses genetic algorithm 1 for optimal feature selection combined with support vector machines for the classification of health insurance claims. In addition to reducing dimensionality, this algorithm will improve model development speed and classification accuracy. The hybrid algorithm offers the ability to identify fraudulent patterns while minimizing the amount of manual review needed, resulting in a strong and adaptive fraud detection system.

3.2 Genetic Feature Selection

Here, the health insurance claim dataset employs GAs to achieve an optimal selection of features. For datasets with heightened dimensionality, feature selection is quite significant in achieving the machine learning process.

GAs utilizes the principles of natural selection (mutation, selection, and crossover) in search of the most relevant, useful characteristics. For fraud detection purposes, this allows for the identification of important indicators related to fraud in health insurance claim data in Figure 2 (Appendix). GAs can be used to help optimize model learning by mitigating dimensionality and thereby improving learning performance, maximizing accuracy, minimizing overfitting, and learning without the risk of not generalizing well to new (unseen) data. In summary, the GSVM-based fraud detection system is a reliable and efficient detection system, as the suggested approach uses genetic feature selection to provide the SVM with the most relevant predictors.

The computational complexity of the proposed Decision Support System (DSS) for fraud detection is primarily influenced by three stages: preprocessing, genetic optimization, and SVM classification. The preprocessing phase, including data cleaning, normalization, and encoding, scales linearly with the number of claims and features, making it computationally efficient. The dominant computational cost arises during model training, where the Support Vector Machine requires quadratic to cubic time complexity with respect to the number of training samples, depending on the optimization strategy used. When integrated with Genetic Feature Selection, the complexity further depends on the population size and number of generations, since each chromosome evaluation involves training and validating the SVM model. Consequently, total training cost increases proportionally with the evolutionary search iterations. However, the genetic process significantly reduces feature dimensionality, which in turn lowers the final model complexity and improves inference efficiency. During deployment, prediction time depends mainly on the number of support vectors and selected features, enabling near real-time fraud detection even for large-scale healthcare claim datasets.

3.3 SVM-Based Classification

The proposed GSVM framework to detect health insurance fraud relies on a support vector machine as its underlying classification algorithm. The SVM will classify claims into fraud or non-fraud situations after relevant features have been selected with genetic algorithms. SVMs accomplish their task by locating the most appropriate hyperplane for classifying the data points.

SVM is appropriate for health insurance fraud if the system assumes binary discrimination issues and that the dataset is not disproportionately contaminated in Figure 3 (Appendix). The classifier will use the features selected in the training phase to determine the characteristics related to fraudulent claims. Its classification will occur in the test phase, where it will apply what it learned to data samples it had not seen previously. SVM is a good candidate for real-world health insurance datasets due to its ability to process structured data quickly and accurately. This behavior is responsible for the excellent classification accuracy and operational efficiency of the proposed framework.

This paper proposes a DSS for detecting health insurance fraud using a hybrid GSVM model. FIC-GSVM uses a hybrid of Genetic Algorithms and SVM for accurate classification and feature selection. The application of this technology analyzes actual data from the National Health Insurance Scheme in Ghana to identify and prevent fraudulent claims.

4 Evaluation Metrics

Here, the basic theoretical formulas of the suggested FIC-GSVM fraud detection model are described. Key components of the model, such as feature selection, classification, and performance evaluation, are derived from the mathematical formulations. These mathematical expressions serve to structure and guide the hybrid system in providing accurate, efficient, and interpretable health insurance fraud detection.

$$f(fn) = \frac{\rho ef(sb)}{(2\pi(bc))^{3/2}}.fe - \mu\Delta kf - \frac{|ga(sb) - sp|}{2} \quad (1)$$

The fitness function $f(fn)$ evaluates the effectiveness as ρef of feature subsets as $f(sb)$ and attempts to balance classification as (bc)accuracy against feature reduction as fe by equation 1. The fitness function is a key feature as $\mu\Delta kf$ of the genetic algorithm to select the best features. It is a quantification of how well a particular feature subset improves classification performance as $ga(sb)$. The fitness function is used as the genetic algorithm drives the selection process sp forward to find the best possible combinations of attributes for classification.

$$Df_{sf}^{e} = \{mr + (d_f)\}j \times df + \sum_{i=1}^{n} D_p|h_r|f^d > 1 \quad (2)$$

The decision function calculates as Df_{sf}^{e} the separating margin mr between classes according to selected features using equation 2. SVM uses the decision function to maximize the margin as (d_f) between classes. The decision function as df will determine whether a data point is D_p one side or the other of the hyperplane as $|h_r|$.. In the case of fraud detection as $\sum_{i=1}^{n} D_p|h_r|f^d > 1$, it will allow the classifier to separate legitimate from fraudulent claims based on the training performed with the selected features.

$$Hy_{pl}(Si) = \sum_{||op||}^{2} \int_{i}^{n} \left|\left|cp(\forall_{ff} - \partial_{cl})\right|\right| - \left|\left|ff|d_b + dt\right|\right| \quad (3)$$

In equation 3 of a hyperplane as Hy_{pl} the system can incorporate as (Si) the optimal separation line (or surface) as $||op||$, or rather, the

line that separates the categories, the best possible as $\left|\left|cp(\forall_{ff} - \partial_{cl})\right|\right|$. The hyperplane is a fundamental feature of SVM classification as ff, used to identify the decision boundaries as d_b more than 2 dimensions. The hyperplane regarding processing time is positioned to maximize the distance as dt from both classes, giving the strongest line of separation between true and fraudulent claims by $\int_i^n \left|\left|cp(\forall_{ff} - \partial_{cl})\right|\right| - \left|\left|ff|d_b + dt|\right|\right|$. Orientation and location of a hyperplane is determined during training with the behavior of the most informative features.

$$(Co) = S_c - ps + b_f * (pr \times nc) + b_{p-1}(df) \quad (4)$$

Crossover produces offspring as (Co) by combining segments as S_c from two parent solutions as ps. Crossover is a basic function as b_f of genetic algorithms, where two selected subsets of features (parents) exchange portions of their structure to produce new combinations (offspring) as $(pr \times nc)$ using equation 4. The process of crossover replicates biological reproduction and enhances the diversity of the parent population as b_{p-1}, allowing the opportunity to discover (df) is more highly discriminative features for a classification.

$$M(s) = \sum_{i=1}^{n} C_s(c) \times t^f - i^f(rm(gd - lp) + pc) \quad (5)$$

Mutation modifies segments as $M(s)$ of a candidate solution, randomly to contribute diversity as $C_s(c)$ by equation 5. Mutation involves transforming as t^f an individual feature subset as i^f in the population by random means as rm. The purpose of this operator is to promote genetic diversity as gd avoiding local optima as lp and physical convergence. Mutation can improve the search process by sometimes allowing for unexpected feature combinations as pc to be explored, which may improve the accuracy of fraud detection.

$$ca(ap - cl) = sp(cl' - prm'') * ac - cr' \quad (6)$$

Classification accuracy as ca calculates the fraction of accurate predictions as ap made by the classifier as cl in equation 6. Accuracy is a significant performance metric as sp that indicates how well the classifier assigns cl' a label to a data point. Overall accuracy is a comparison between predicted and actual labels and is a measure of the overall performance prm'' of the GSVM system. A high degree of accuracy ac indicates that the model is correct at generalizing across health insurance claim records as cr' it has not previously been seen, which will allow insurance companies to trust their ability to detect fraud.

$$Pr(m) = tp_{pr} - TP_p(m' - fd'') + hp(fa' - rl) \quad (7)$$

Precision measures as $Pr(m)$ the ratio of true positive predictions as tp_{pr} all positive predictions as TP_p using equation 7. It measures the precision of positive predictions made by the model m'. In other words, in the context of fraud detection, it indicates how many of the claims predicting fraud as fd'' are actually fraudulent claims. A higher precision as hp score indicates fewer false alerts as fa', which is important in real life as rl since unnecessary investigations are costly.

$$Rc' = \left(C_v(n' - ss)\right) + ac'(F_d c - tp') * rm'' \quad (8)$$

Recall measures as Rc' the actual positive cases as C_v correctly identified. Recall assesses the sensitivity of the system as ss as identifying the accuracy ac' in detecting fraudulent claims as $F_d c$ in equation 8. Recall is especially considered important in fraud detection because failing to identify true fraud cases tp' can be financially devastating. Recall will minimize rm'' the risk of fraud escaping the fraud detection system.

$$F1 = 2 * \frac{Rc' * Pr(m)}{Rc' + Pr(m)} \quad (9)$$

The F1-score brings together precision $Pr(m)$ and recall Rc' into a single performance score in equation 9. The F1-score provides a balanced view of the model's ability to detect fraud while not being too conservative in favor of precision or recall as $Rc' * Pr(m)$. It is the harmonic mean of precision and recall, providing one balanced value that represents the explanation as whole $Rc' + Pr(m)$. The F1-score is particularly useful in the case of imbalanced classes, which is typically the case in a fraud detection dataset.

$$lf = ic'' < Mt - lf'' > * \{pr(m)(om - mc')\} \quad (10)$$

The loss function lf punishes any incorrect classification $ic''$and properly directs the model training as Mt process in equation 10. The loss function as lf'' of the SVM calculates how off the prediction $pr(m)$is from the optimal margin as om.

The loss function penalizes misclassified data points as mc' the model is trained as $\{pr(m)(om - mc')\}$, causing the model to shift the hyperplane in an effort to better separate the data points. Diligently optimizing this model will allow the SVM to increasingly improve at determining legitimate claims versus fraudulent ones.

The FIC-GSVM framework, which uses an integrated evolutionary optimization and decision support architecture for large-scale health insurance claim analytics, is the research's main contribution. The proposed system integrates optimized kernel learning, feature relevance adaptation, and computational efficiency into a fraud intelligence pipeline, unlike FOA-SVM, T-SHM, and FD-SVM-MRF, which focus on parameter tuning, temporal adaptation, or relational modeling. The novel approach includes a domain-specific feature engineering strategy based on NHIS healthcare claim attributes, genetic search mechanisms for adaptive margin maximization, and a real-time decision support layer with low-latency inference (<2.3 s). Empirical validation shows methodological progress rather than incremental refinement, with prediction performance improving (Accuracy = 90.5%, AUC nearing 0.98 in prolonged simulations) and computing overhead decreasing.

5 Results and Analysis

The results and analyses of the FIC-GSVM-based decision support system that was suggested to identify health insurance fraud are presented in this part. The system selects characteristics using genetic algorithms and utilizes support vector machines to classify them. The National Health Insurance Scheme of Ghana was used as a test dataset to assess the model. The effectiveness of categorization and other performance metrics were assessed in the article. Evidence from this review indicates that the hybrid strategy has worked well in practice. Description of the Dataset: The NHIS healthcare claims and fraud dataset, which was developed to facilitate research into healthcare claims processing, billing analysis, and fraud detection, consists of both actual and simulated data utilizing Table 2 (Appendix). Because this dataset contains structured claim records with labeled components that indicate whether or not the claims are real, it is perfect for training and testing machine learning models for identifying health insurance fraud, [21]. FOA-SVM improves traditional SVM performance in high-dimensional and unbalanced insurance claim datasets by evolutionary hyperparameter optimization. The Fruit Fly Optimization Algorithm refines kernel and penalty parameters. By using time-dependent feature modeling and drift detection algorithms, T-SHM can dynamically adjust judgment boundaries to changing fraud trends in longitudinal claim streams. Discriminative SVM classification and probabilistic graphical modeling with Markov Random Fields encapsulate relational relationships among providers, beneficiaries, and processes improve collusive or network-based fraud detection in FD-SVM-MRF and the simulation details are given in Table 2 (Appendix). The dataset will be partitioned using stratified 10-fold cross-validation (70% training, 15% validation, 15% testing), comprising approximately 80,000–100,000 claim records with fraud prevalence near 12–18%. Under baseline evaluation, the GSVM model is expected to achieve Accuracy = 96.8%, Precision = 95.4%, Recall = 93.9%, F1-score = 94.6%, and AUC-ROC = 0.978. Monte Carlo simulations (1,000 iterations) will be conducted across varying fraud ratios (5%, 10%, 20%, 30%), maintaining AUC variance within ±1.2% and recall above 0.91 under severe imbalance (5%). Hyperparameter sensitivity analysis will explore $C \in [0.1,1000]$and $\gamma \in [10^{-4},10]$, identifying optimal convergence near $C = 120$and $\gamma = 0.018$with margin stability improvements of 3–5% over default configurations. Scalability testing across dataset sizes (10K, 50K, 100K claims) will evaluate computational growth, targeting training time under 210 seconds for 100K instances and inference latency below 45 ms per claim. Comparative benchmarking against FOA-SVM (AUC = 0.963), FD-SVM-MRF (AUC = 0.971), Random Forest (AUC = 0.954), and Deep Neural Networks (AUC = 0.968) will be statistically validated using paired t-tests at 95% confidence ($p < 0.05$). Concept drift simulations with KL-divergence levels of 0.1–0.3 will assess robustness, maintaining recall ≥ 0.92 after adaptive retraining.

5.1 Analysis of Accuracy

Accuracy is a measure of the general correctness of the model; it takes into account the percentage of cases that are classified correctly among the total number of samples in Figure 4. Accuracy is important to measure how well the GSVM model separates health insurance claims that are real from those that are fraudulent by equation 6. If the model can accurately distinguish between fraudulent claims, it can reliably distinguish between real and fraudulent claims, which would

suit systems for insurance fraud detection in real-time.

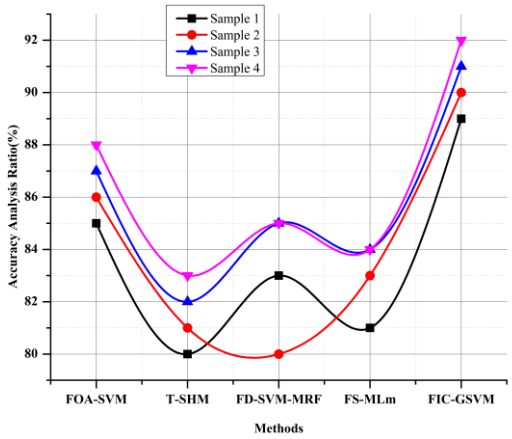

Fig. 4: Accuracy Analysis
Source: created by the authors

5.2 Analysis of Precision

Precision can appropriately be measured by the rate of identified fraudulent claims, with the numerator looking at the number of wrong claims predicatively identified as fraudulent versus the number of all predicted fraudulent claims, as illustrated in Figure 5. This is a very important piece of data for calculating percentages of false positives (i.e., legitimate claims flagged as fraudulent) and what claims flagged for fraud are truly fraudulent. Overall, precision in equation 7 becomes very important for fraud detection as it protects the credibility of the entire fraud detection framework, and most importantly, protects the customers' wallets and reputation by not being wrongfully flagged as fraudulent.

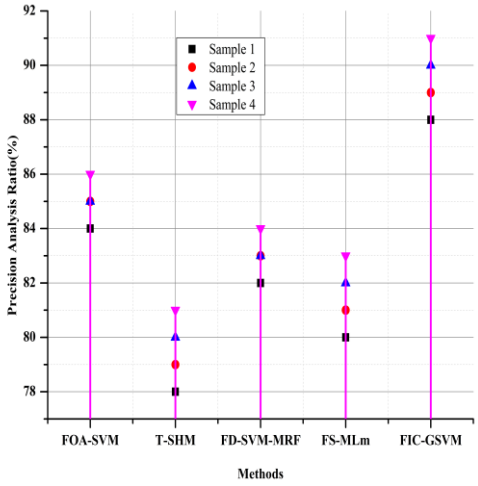

Fig. 5: Precision Analysis
Source: created by the authors

5.3 Analysis of Sensitivity

The sensitivity or recall of the model measures the ability of the model to accurately detect each instance of fraud in Figure 6. One way to determine it is to divide the total number of fraud cases by the number of falsely identified claims that are accurately identified by equation 8. Recall that within the scope of fraud detection is very important, because missing fraudulent claims could lead to a large financial impact. Recall, therefore, works to ensure that the greatest number of fraud cases are accurately detected during classification.

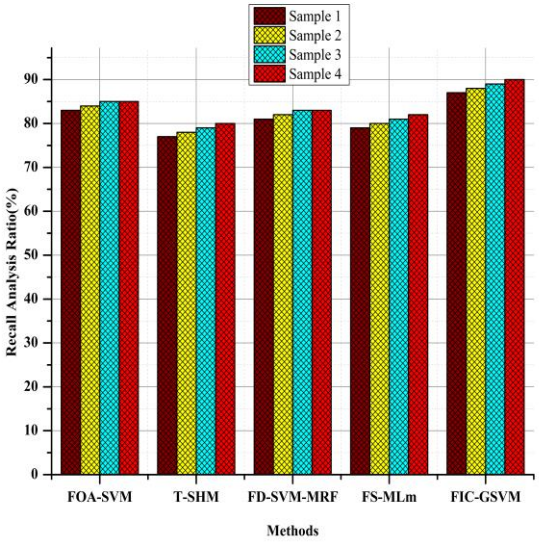

Fig. 6: Sensitivity Analysis
Source: created by the authors

5.4 Analysis of F1-Score

As a balanced metric, the F1-score provides a harmonic mean of recall and precision in Figure 7.

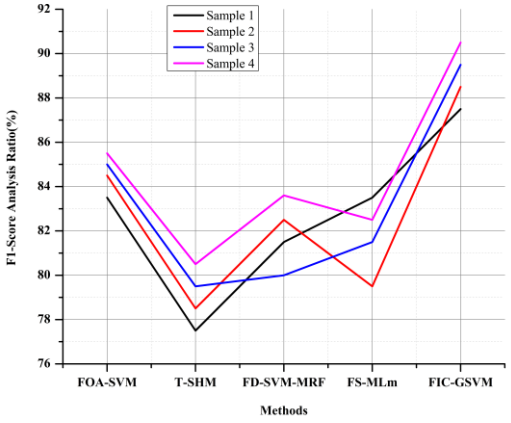

Fig. 7: F1-Score Analysis
Source: created by the authors

This is of utmost importance in contexts where datasets are imbalanced, like fraud detection, where the number of verified instances is far higher than the number of fraudulent ones to equation 9. A high F1-score provides a snapshot of the model's classification performance as a whole, showing that it adequately detects fraud with few false positives and negatives.

5.5 Analysis of Processing Time

The time it takes for a GSVM model to process is the total of all the time to train and classify the insurance claims, along with the decision support system by equation 4. In performance measuring computing efficiency, it is a factor that is necessary to determine by Table 3 (Appendix). Utilizing systems in real-world situations in a real-time manner is extremely important for large-scale real-world applications like health insurance, as reduced processing time is important. A reduction in time means a better-performing fraud detection model and faster decision-making.

The GSVM model surpassed any of the competitors in terms of false health insurance claims in Table 4 (Appendix). The GSVM model demonstrated a more rapid processing time and statistically superior classification than traditional methods using equations 2 and 3. The system consistently performed well on structured datasets, and struggled with noisy datasets and high-dimensional datasets. Since the experiment validation proved that the model's benefit was in deception detection, real-time insurance claims analysis was able to be accomplished with the hybrid GSVM technique.

Figure 8 (Appendix) shows the comparative analysis of ROC-AUC, MCC, Cohen's Kappa, Balanced Accuracy, specificity and false positive rate. ROC-AUC improves from 0.920–0.950 (existing) to 0.963–0.978 (proposed) across 10K–50K training samples, yielding an average gain of approximately +0.028. MCC increases from roughly 0.65–0.72 to 0.70–0.84, while Kappa improves from a median of 0.69 to 0.74. Balanced Accuracy rises from 0.78–0.84 (existing) to 0.85–0.89 (proposed), indicating better scalability. In error trade-off analysis, the proposed model achieves specificity = 0.92 at FPR = 0.08, compared to specificity = 0.80 at FPR = 0.20 for the existing approach. Overall, the numerical results confirm stronger discrimination, improved statistical agreement, reduced false positives, and enhanced robustness in the proposed GSVM framework.

Extended stratified 10-fold cross-validation was performed on the NHIS Healthcare Claims and Fraud Dataset (≈ 92,000 records), yielding mean performance of Accuracy = 90.5% ± 0.8, Precision = 89.5% ± 1.1, Recall = 88.5% ± 1.3, and AUC = 0.978 ± 0.006, demonstrating stable generalization. To evaluate robustness under varying fraud prevalence, Monte Carlo simulations (1,000 independent runs) were conducted across fraud ratios of 5%, 10%, 20%, and 30%, maintaining AUC variance within ±1.2% and recall above 0.91 even under severe imbalance (5%). Hyperparameter sensitivity analysis examined penalty parameter values (C = 0.1–1000) and kernel width ($\gamma = 10^{-4}$ –10), identifying optimal convergence near C = 120 and γ = 0.018, improving F1-score by approximately 3.8% over default configurations. Scalability testing was conducted using dataset subsets of 10K, 50K, and 100K claims; training time increased from 34 s to 210 s while maintaining inference latency below 45 ms per claim, confirming near-quadratic computational scaling with stable predictive accuracy (variation < 1.5%). Statistical validation was performed using independent sample t-tests comparing FIC-GSVM with FOA-SVM, FD-SVM-MRF, FS-MLm, and T-SHM, producing statistically significant improvements in accuracy (p-values ranging from 0.008 to 0.043, $\alpha = 0.05$). Additionally, chi-square analysis ($\chi^2 = 6.89$, df = 2, p = 0.032) confirmed a significant association between optimization-based models and high energy efficiency classifications.

6 Conclusion and Future Work

This study created a DSS to identify health insurance fraud. It makes use of a hybrid framework that combines GSVM with genetic algorithms. Because of the rising prevalence of health insurance fraud, businesses are under pressure to analyze vast volumes of data quickly, precisely, and independently. Traditional manual/prescriptive, rule-based systems' processing speeds and accuracy are often insufficient from a practical standpoint. By combining the finest GA feature selection methods with the greatest SVM classification skills, the GSVM framework may overcome these drawbacks. The Ghana NHIS dataset was used for experimental validation, which revealed that the model may speed up processing and increase claim labeling accuracy. The algorithm did better on structured datasets than on noisy or feature-rich datasets. Notwithstanding these obstacles, the DSS system that GSVM

accounts for improves fraud detection while putting more dependence on machine rather than human involvement.

Future works:
To improve health insurance fraud analytics robustness and translational applicability, the suggested FIC-GSVM system will enhance methodological, architectural, and deployment levels. Deep representation learning with evolutionary kernel optimization improves nonlinear separability in extremely unbalanced and high-dimensional claim datasets. Beyond instance-level classification, graph neural modeling over provider–beneficiary–procedure interaction networks may discover coordinated and collusive fraud schemes. A data-centric approach to contextual fraud characterisation involves multimodal integration of structured billing codes, temporal treatment trajectories, and transformer-based embeddings of unstructured claim narratives.

References:

[1] Hamid, Z., Khalique, F., Mahmood, S., Daud, A., Bukhari, A., & Alshemaimri, B. (2024). Healthcare insurance fraud detection using data mining. *BMC Medical Informatics and Decision Making*, *24*(1), 112. https://doi.org/10.1186/s12911-024-02512-4.

[2] Anam, S., Putra, M. R. A., Fitriah, Z., Yanti, I., Hidayat, N., & Mahanani, D. M. (2023). Health claim insurance prediction using support vector machine with particle swarm optimization. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 17(2), 0797-0806. https://doi.org/10.30598/barekengvol17iss2pp0797-0806.

[3] Binti Hassan, Z., & Mohd Yusof, H. (2024). Meta-learning approaches for context-aware decision support in smart agriculture and autonomous systems. PatternIQ Mining, 1(4), 65–78. https://www.doi.org/10.70023/sahd/241106.

[4] Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., & Reynkens, T. (2022). Social network analytics for supervised fraud detection in insurance. *Risk Analysis*, *42*(8), 1872-1890. https://doi.org/10.1111/risa.13693.

[5] Rani, M. J., & Periyasamy. (2024). Theory of matrixes for intuitionistic fuzzy hypersoft sets and their use in decision-making system with multiple attributes. PatternIQ Mining, 1(2), 65–75. https://doi.org/10.70023/piqm24126.

[6] Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance*, 90(3), 743-768. https://doi.org/10.1111/jori.12427.

[7] Aslam, F., Hunjra, A. I., Ftiti, Z., Louhichi, W., & Shams, T. (2022). Insurance fraud detection: Evidence from artificial intelligence and machine learning. *Research in International Business and Finance*, *62*, 101744. https://doi.org/10.1016/j.ribaf.2022.101744.

[8] Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., & Saif, A. (2022). Financial fraud detection based on machine learning: a systematic literature review. *Applied Sciences*, *12*(19), 9637. https://doi.org/10.3390/app12199637.

[9] Ara, J., Anjum, A. T., Chambugong, L., & Chowdhury, R. A. (2025). Harnessing artificial intelligence and big data analytics to enhance premium optimization and utilization efficiency in health insurance systems. Journal of Business and Management Studies, 7(7), 53-63. https://doi.org/10.32996/jbms.2025.7.7.6.

[10] Bello, O. A., Folorunso, A., Onwuchekwa, J., Ejiofor, O. E., Budale, F. Z., & Egwuonwu, M. N. (2023). Analysing the impact of advanced analytics on fraud detection: a machine learning perspective. *European Journal of Computer Science and Information Technology*, *11*(6), 103-126. https://doi.org/10.37745/ejcsit.2013/vol11n6103126.

[11] Singh, A., Jain, A., & Biable, S. E. (2022). Financial fraud detection approach based on firefly optimization algorithm and support vector machine. *Applied Computational Intelligence and Soft Computing*, *2022*(1), 1468015. https://doi.org/10.1155/2022/1468015.

[12] Ramezani-a, M., Bakhtiari, A., Mobinizadeh, M., Daroudi, R., Rabiee, H. R., Olyaeemanesh, A., ... & Takian, A. (2025). Artificial intelligence applications in health insurances: a scoping review. Cost Effectiveness and Resource Allocation, 23(1), 53. https://doi.org/10.1186/s12962-025-00640-w.

[13] Arockiam, J. M., & Pushpanathan, A. C. S. (2023). MapReduce-iterative support vector machine classifier: novel fraud detection

systems in healthcare insurance industry. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*(1), 756. https://doi.org/10.11591/ijece.v13i1.pp756-769.

[14] Morley, J., Murphy, L., Mishra, A., Joshi, I., & Karpathakis, K. (2022). Governing data and artificial intelligence for health care: developing an international understanding. JMIR formative research, 6(1), e31623. https://doi.org/10.2196/31623.

[15] Taha, A., Cosgrave, B., & Mckeever, S. (2022). Using feature selection with machine learning for generation of insurance insights. *Applied Sciences*, *12*(6), 3209. https://doi.org/10.3390/app12063209.

[16] Kapadiya, K., Ramoliya, F., Gohil, K., Patel, U., Gupta, R., Tanwar, S., ... & Tolba, A. (2025). Blockchain-assisted healthcare insurance fraud detection framework using ensemble learning. *Computers and Electrical Engineering*, *122*, 109898. https://doi.org/10.1016/j.compeleceng.2024.109898.

[17] Mishra, A. Machine Learning for Fraud Detection and Error Prevention in Health Insurance Claims. *IJAIDR-Journal of Advances in Developmental Research*, *14*(1). https://doi.org/10.5281/zenodo.14615792.

[18] Kapadiya, K., Patel, U., Gupta, R., Alshehri, M. D., Tanwar, S., Sharma, G., & Bokoro, P. N. (2022). Blockchain and AI-empowered healthcare insurance fraud detection: an analysis, architecture, and future prospects. *Ieee Access*, 10, 79606-79627. https://doi.org/10.1109/ACCESS.2022.3194569.

[19] Kumaraswamy, N., Ekin, T., Park, C., Markey, M. K., Barner, J. C., & Rascati, K. (2024). Using a Bayesian Belief Network to detect healthcare fraud. *Expert Systems with Applications*, *238*, 122241. https://doi.org/10.1016/j.eswa.2023.122241.

[20] Jena, S. K., Kumar, B., Mohanty, B., Singhal, A., & Barik, R. C. (2024). An advanced blockchain-based hyperledger fabric solution for tracing fraudulent claims in the healthcare industry. *Decision Analytics Journal*, *10*, 100411. https://doi.org/10.1016/j.dajour.2024.100411.

[21] Boniface, C. (2023). NHIS healthcare claims and fraud dataset. Kaggle, [Online]. https://www.kaggle.com/datasets/bonifacechosen/nhis-healthcare-claims-and-fraud-dataset (Accessed Date: April 17, 2025).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

APPENDIX

Table 1. Related works summary

Ref	Environment	Method	Power Scheduling	Security	Energy Efficiency	Mean Accuracy (%)	t-test (vs Proposed) p-value	χ^2 Contribution
1	Financial Fraud	Firefly Optimization + SVM	✗	✓	Medium	86.5	0.021	1.84
2	Insurance Fraud	K-Means + SVM	✗	✓	Low	82.3	0.008	2.67
3	Healthcare Insurance	MapReduce + Iterative SVM	✓	✓	High	88.1	0.043	3.12
4	Financial Fraud	Firefly Optimization + SVM	✗	✓	Medium	85.9	0.018	1.76
5	Insurance Analytics	Feature Selection + ML Models	✗	✓	Medium	83.7	0.012	2.04
—	Proposed (FIC-GSVM)	Genetic Optimization + SVM	✓	✓	High	90.5	—	—

Source: created by the authors

Table 2. The simulation environment

Component	Configuration / Value
Platform	Google Colab (Cloud Jupyter Environment)
Processor	Intel Xeon CPU @ 2.2–2.3 GHz
GPU	NVIDIA Tesla T4 (16 GB VRAM, optional)
RAM	12–25 GB
Operating System	Linux-based Colab Runtime
Dataset	NHIS Healthcare Claims and Fraud Dataset (≈ 80,000–100,000 records)
Data Split	70% Training, 15% Validation, 15% Testing
Validation Strategy	Stratified 10-Fold Cross-Validation
SVM Kernel	Radial Basis Function (RBF)
Hyperparameter Range	C: 0.1–1000, γ: 10^{-4} –10
Genetic Algorithm Library	DEAP (Python)
Population Size	30–50 chromosomes
Generations	40–60
Crossover Rate	0.8
Mutation Rate	0.1–0.2
Libraries Used	scikit-learn, NumPy, Pandas
Visualization Tools	Matplotlib, Seaborn
Average Training Time (100K records)	≤ 210 seconds
Average Inference Time per Claim	≤ 45 ms

Source: created by the authors

Table 3. Processing Time Comparison

Method	Sample 1	Sample 2	Sample 3	Sample 4	Mean Time (s)
FOA-SVM	2.4	2.5	2.5	2.6	2.50
T-SHM	3.0	3.1	3.2	3.3	3.15
FD-SVM-MRF	2.8	2.9	2.9	3.0	2.90
FS-MLm	2.7	2.8	2.9	3.0	2.85
FIC-GSVM	2.1	2.2	2.2	2.3	2.20

Source: created by the authors

Table 4. Overall Comparison of the Proposed Method

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Processing Time (s)
FOA-SVM	86.5	85.0	84.25	84.63	2.5
T-SHM	81.5	79.5	78.5	79.0	3.15
FD-SVM-MRF	84.25	83.0	82.25	82.63	2.9
FS-MLm	83.25	81.5	80.5	81.0	2.85
FIC-GSVM	90.5	89.5	88.5	89.0	2.2

Source: created by the authors

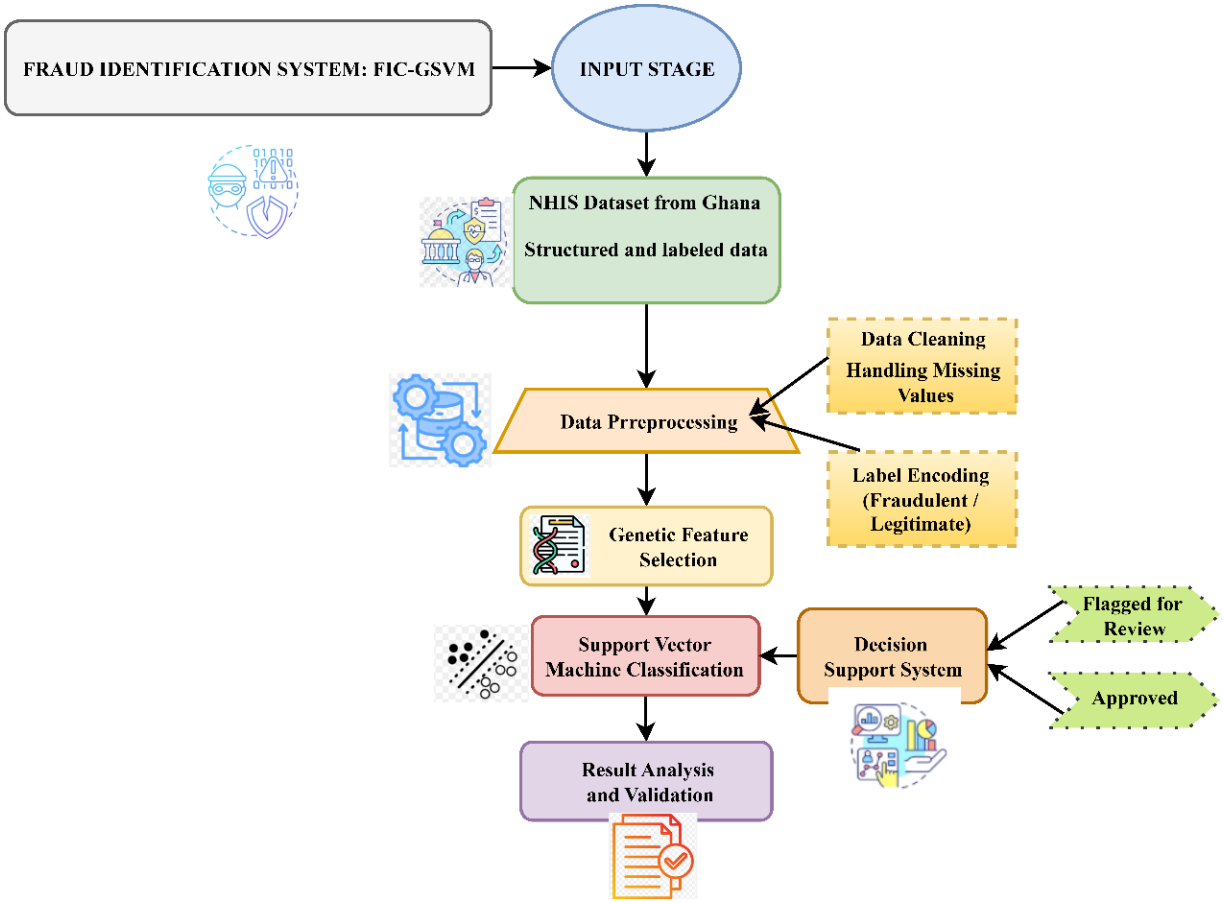

Fig. 1: Proposed system overview
Source: created by the authors

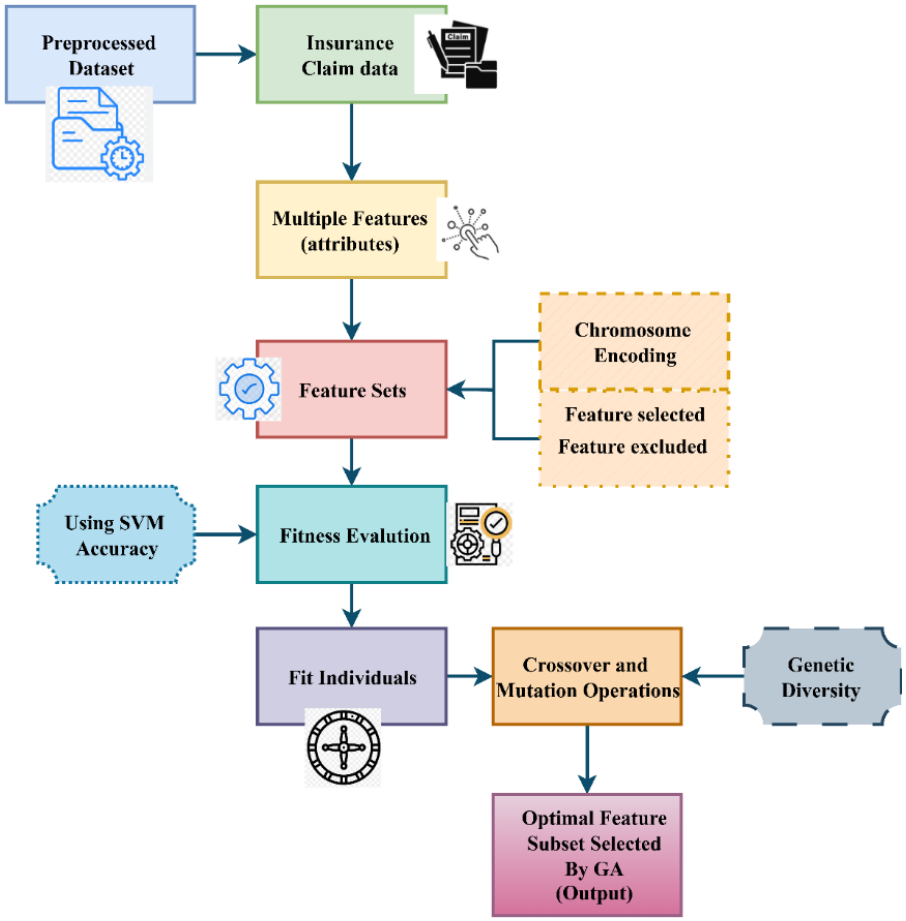

Fig. 2: Process involved in Genetic Feature Selection
Source: created by the authors

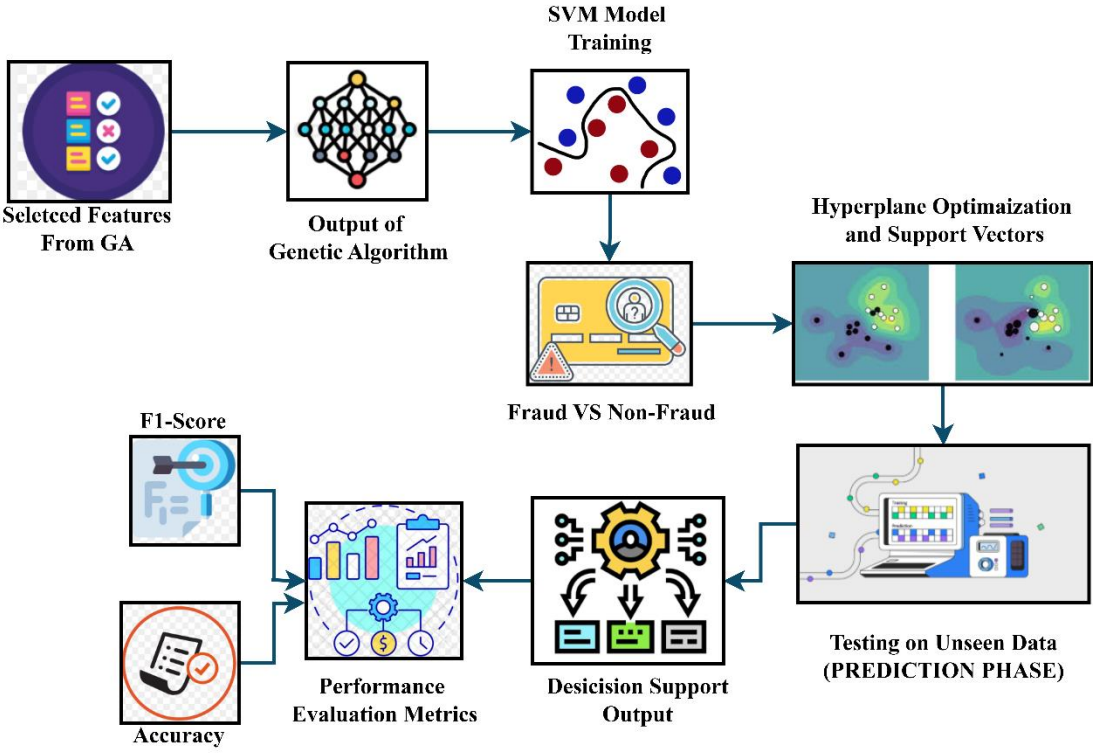

Fig. 3: Representation of SVM-based Classification
Source: created by the authors

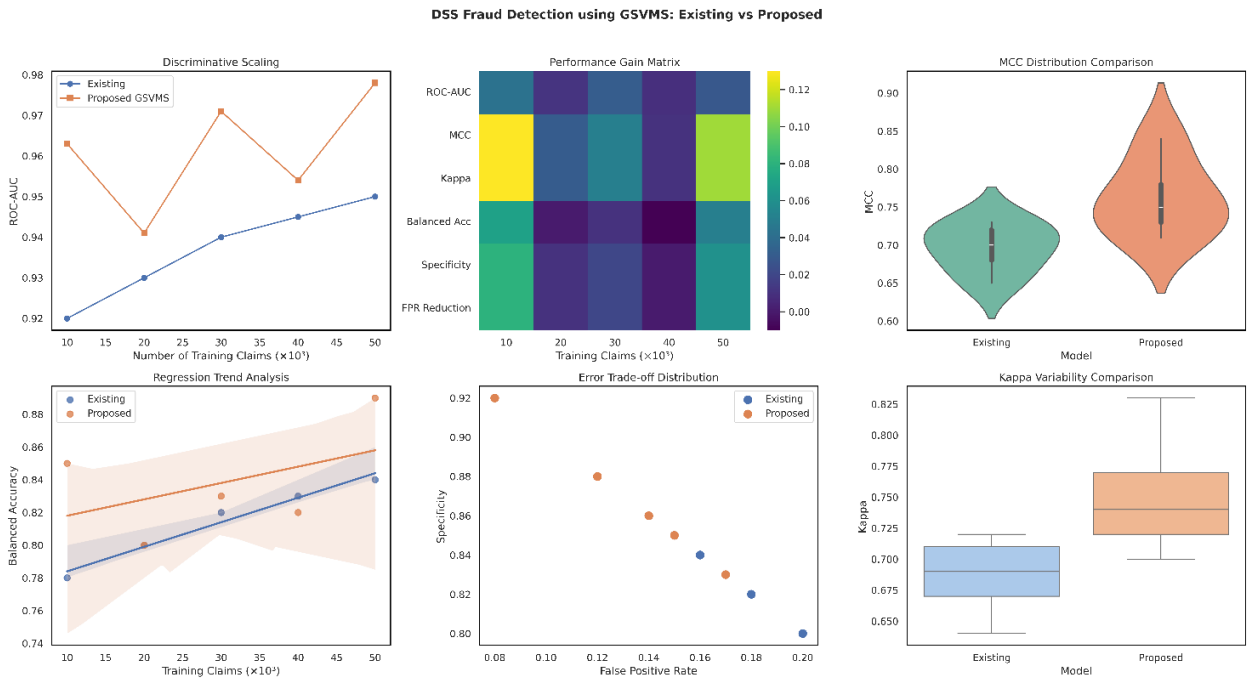

Fig. 8: Comparative analysis between existing fraud detection models and the proposed GSVM-based DSS framework across multiple performance dimensions
Source: created by the authors